

Illustrating and Researching the Common European Framework Levels with a Multilingual Online Platform

Katrin Wisniewski, Technical University Dresden
 Katrin.Wisniewski@tu-dresden.de

INTRODUCTION to MERLIN (Multilingual Platform for the European Reference Levels: Exploring Interlanguage in Context)

- **background:** widespread use of the CEFR, but not enough empirical data illustrating the meaning of CEFR levels
- **aim:** MERLIN will be a freely accessible trilingual (CZECH; GERMAN; ITALIAN) online platform that serves for the illustration of CEFR levels. It helps teachers, test developers, teacher trainers, and textbook authors to get a concrete picture of CEFR levels by making accessible and searchable a large number of full-text learner productions. It is based on a richly annotated learner corpus.

DATA PREPARATION

DATA

- taken from standardized language tests (developed by telc for Italian & German, by ÚJOP for Czech)
- written production tasks (3-4 different tasks per level/language)

	GERMAN	ITALIAN	CZECH
A1	229	229	
A2	228	224	131
B1	231	223	176
B2	225	222	131
C1	226		

Table 1: number of tests in MERLIN corpus

RE-RATINGS

- professional raters
- rater training
- MERLIN analytical rating grid + holistic CEFR scale
- rating criteria: grammatical accuracy, vocabulary range & control, sociolinguistic appropriateness, coherence & cohesion, orthography control
- analysis of quality of ratings (reliability measures; MFRM)

TRANSCRIPTION

- transcription editor (xml-based, xml mind ©EURAC)
- transcription tagset (characteristics of handwritten texts like corrections, deletions, emoticons were kept ...)
- anonymization
- reliability checks

ANNOTATION STAGE

ANNOTATED FEATURES

ERRORS

OTHER FEATURES OF LEARNER LANGUAGE

TARGET HYPOTHESES

FALKO*

SOURCE	METHODOLOGY	EXEMPLARY TAGS
CEFR scales	operationalisation of chapter 5 scales (vocabulary range & control, grammatical accuracy, orthography, coherence & cohesion, sociolinguistic appropriateness)	content jumps greetings repetitions intelligibility
users (teachers, testers, learners)	<ul style="list-style-type: none"> needs analysis (questionnaires & expert interviews) textbook analyses language test analyses 	verb mood, verb aspect incorrect use of prepositions capitalization errors
research	analyses of research literature - learner corpora, SLA, LT (language-specific & overarching)	REQUESTS connectors, text-structural means, reference problems ... word order, inflection ... semantic/formal lexical errors ... collocations, idioms ... single variational aspects like double pronouns (ITA)
learner texts	inductive qualitative analyses of 10 texts per level & language	formality problems ITA: lexicalised clitics

THE ANNOTATION PROCEDURE

PREPARATION

(e.g. tokenization, transcription checks)

PILOTING

(annotation guidelines & schemes)

MANUAL ANNOTATION I: GRAMMAR & ORTHOGRAPHY TARGET HYPOTHESES I

AUTOMATIC ANALYSES (e.g. POS; parser...)

MANUAL ANNOTATION II: VOCABULARY, SOCIOLOGICALISTICS, PRAGMATICS TARGET HYPOTHESES II

TOOLS/FORMATS

- reuse & development of new NLP components
- PAULA exchange format
- MMAX2 (manual annotation of phenomena)
- EXMARALDA & Excel with Falko Add-ins (TH)
- relANNIS (visualisation)

FINAL STAGE

INTERFACE DESIGN, PILOTING, REVISION

- preparation of platform (user-based, didactically oriented specification of interface functionalities)
- piloting with target groups
- revision stage

THE MERLIN PLATFORM

- contains a richly annotated trilingual learner corpus (Czech, German, Italian)
- offers free online access, open source license for tools & resources
- shows full texts & language test tasks
- shows individual CEFR profile on rating grid
- is searchable for a wide spectrum of L2-related phenomena
- can sort texts according to CEFR level – tasks – linguistic phenomena
- produces word lists (e.g. showing words of a particular word class according to CEFR levels)
- delivers selected statistical measures (e.g. vocabulary indices, complexity & accuracy measures)
- exports results, creates ad hoc worksheets
- compares features of learner languages from 3 language families (Roman, Slavic, Germanic)

PROJECT INFORMATION

MERLIN is an LLP project and runs from 01/2012 to 12/2014.
 Coordinated by Institute for Romance Studies, TU Dresden; by Katrin Wisniewski (parental leave) and Karin Schöne
 Partners: EURAC Bolzano, Eberhard-Karls-University Tübingen, telc Frankfurt, bfi Oberösterreich, Charles University Prague
 Contact: Katrin.Wisniewski@tu-dresden.de, Karin.Schoene@tu-dresden.de
 This project has been funded with support from the European Commission (518989-LLP-2011-DE-KA2-KA2MP). This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

* MERLIN cooperates with the FALKO project, run by Anke Lüdeling: <http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko>



Education and Culture DG

Lifelong Learning Programme